

Contigi csb10a_v1:

pliki ze strony <http://csgenome.sggw.pl/en-us/resources/>

Fragmenty plików:

- Gene_Annotation_v1.xlsx :

#	A	B	C	D	E	F	G	H	I	J	K
1	Lp	Gene	Start_Gene	Stop_Gene	Strand	ACC_NO	Species	Protein Product	goBP	goMF	goCC
2	3	gene_1#CSB10A_v1_contig_3	1043	5674	+	XP_002323740	Populus trichocarpa	histone acetyltransferase		0 nucleic acid binding	0
3	4	gene_1#CSB10A_v1_contig_4	3352	16648	+	XP_002323740	Populus trichocarpa	histone acetyltransferase		0 nucleic acid binding	0
4	9	gene_2#CSB10A_v1_contig_5	13220	partial	+	XP_002523117	Ricinus communis	galacturan 1,4-alpha-galacturonidase, pu	carbohydrate n	polygalacturonase ac	0
5	12	gene_1#CSB10A_v1_contig_8	9819	14148	+	BAH57245	Arabidopsis thaliana			0	0
6	15	gene_4#CSB10A_v1_contig_8	35574	33792	-	BAF07113	Oryza sativa Japonica Group	Spermatophyta; Magnoliophyta; Lil		0	0
7	18	gene_2#CSB10A_v1_contig_10	9561	13870	+	NP_193037	Arabidopsis thaliana	oxidoreductase, zinc-binding dehydrogen	metabolic proc catalytic activity bindi	vacuole pi	0
8	19	gene_3#CSB10A_v1_contig_10	23201	24610	+	ABA81857	Solanum tuberosum	ripening regulated protein-like	metabolic proc catalytic activity bindi		0
9	20	gene_4#CSB10A_v1_contig_10	26895	25065	-	AAF82612	Prunus dulcis	self-incompatibility associated ribonuclea	0 RNA binding ribonuc		0
10	23	gene_2#CSB10A_v1_contig_12	2302	9209	+	CAB95025	Nicotiana tabacum	pectin methylesterase	cell wall modifi enzyme inhibitor activ	cell wall	0
11	25	gene_4#CSB10A_v1_contig_12	20046	20702	+	XP_002527622	Ricinus communis	anthranilate 5-aramatic acidtransferase		0 aryltransferase activi	0

- csb10a_v1_predictions.fasta:

```
>gene_2#CSB10A_v1_contig_10146
MRRHFRDWWLGAGFDVPNYFSPIFFQIGNSATGFFIIFALIAAVA...
>gene_1#CSB10A_v1_contig_10147
MAIEIQLINLGMQWPELLKLLSSSFGAERNSALLTYFFLSTMPELYT
```

- csb10a_v1_upstream_1000:

```
>gene_1#CSB10A_v1_contig_3
AAGGACCTTCTCATTCCAATTCTTTTGAATATAAGCCCTTGAGCTTTTCCCTAAAC...
>gene_1#CSB10A_v1_contig_4
GACTGATTATATATTGGTCCGTTCTGCAAAGGGGAAGCTTTCCCTGAGACGTGTT...
```

-csb10a_v1_contig.fsa:

```
>CSB10A_v1_contig_1
AAGATGAGGTGTTTATAGGACTTTAGTTTTCAAAAATTGAAATTTTAAAATGAGAT...
>CSB10A_v1_contig_4
GTTCAATCGGTTGCTTGGCTAACGGAAAAACAAGAGGTCGGACGAAGA...
```

- GFF_to_Predictions_ID:

ContigID	Model	Feature	Start	Stop	Strand	-	GFF_GeneID	Predictions_GeneID
CSB10A_v1_contig_1	GeneMark.hmm	stop_codon	11744	11746	-	0	1_g	gene_1#CSB10A_v1_contig_1
CSB10A_v1_contig_1	GeneMark.hmm	CDS	11744	11760	-	1	1_g	gene_1#CSB10A_v1_contig_1
CSB10A_v1_contig_1	GeneMark.hmm	CDS	11924	11942	-	0	1_g	gene_1#CSB10A_v1_contig_1
CSB10A_v1_contig_1	GeneMark.hmm	start_codon	11940	11942	-	0	1_g	gene_1#CSB10A_v1_contig_1

1. posiadamy sekwencje contigów (plik csb10a_v1_contig.fsa) => posiadamy długości contigów
2. znamy położenie genów względem contigów (kolumny Start_Gene i Stop_Gene w pliku Gene_Annotation_v1.xlsx)
- ponieważ pkt1 i pkt2 => znamy również sekwencje genów
3. posiadamy drugie źródło sekwencji genów (plik csb10a_v1_upstream_1000.fasta), w którym jest więcej genów niż w danych z pkt2. Dla dodatkowych genów nie mamy pozycji Start_Gene i Stop_Gene.
4. Jak interpretować słowo 'partial' w niektórych kolumnach (Start_Gene i Stop_Gene w pliku Gene_Annotation_v1.xlsx) ?
5. Jak interpretować dane gdy Start_Gene > Stop_Gene (np. gene_2#CSB10A_v1_contig_10146 Gene_Annotation_v1.xlsx) ?

6. Jakie informacje zawiera plik csb10a_v1_predictions.fasta ?
7. Nie wiemy na jakich chromosomach i gdzie leżą contigi.
8. Nie znamy długości chromosomów.

Contigi z markerami BAM:

pliki <https://drive.google.com/drive/folders/0B1KskZHvmCFpTEYtYlJVNnJYWk0?usp=sharing>

Fragmenty plików:

- links3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ncbi_ctg_id	lp_oryg	#super_id	num_bases_in_super	num_contigs_in_super	ordinal_contig_id	length_of_contig	estimated_gap_bef	estimated_gap_aft	start	stop			
2	1	1	0	2496462	46	1	0	23626	0	-299		0	23626	2496462
3	2	2	0	2496462	46	2	1	1500	-299	56167		23327	24827	
4	3	3	0	2496462	46	3	2	8044	56167	-6380	56167	80994	89038	
5	4	4	0	2496462	46	4	3	47152	-6380	-3469		82658	129810	
6	5	5	0	2496462	46	5	4	14055	-3469	-300		126341	140396	
7	6	6	0	2496462	46	6	5	1500	-300	51167		140096	141596	
8	7	7	0	2496462	46	7	6	1500	51167	-299	51167	192763	194263	
9	8	8	0	2496462	46	8	7	36453	-299	-14784		193964	230417	
10	9	9	0	2496462	46	9	8	20179	-14784	7822		215633	235812	

- STC_vs_contigs1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	lp	Read_name	Polskie	Chinese	Read_Len	Ctg_ID_NCBI	Ctg_Length	1st_base_ctg	Last_base_ctg	strand	Read_partner_name	Partner_contig	Real_insert_len	cM
2	7535	STC1_Bam_071_J24_M13.f	1	4	780	138	30580	2252	3031	-	STC1_Bam_071	135	121308	
3	7537	STC1_Bam_001_H06_M13.r	1	4	762	138	30580	3162	3923	+	STC1_Bam_001	139	89446	
4	7538	STC1_Bam_078_D01_M13.f	1	4	861	138	30580	3359	4219	+	STC1_Bam_078	1		
5	7541	STC1_Bam_061_C16_M13.f	1	4	802	138	30580	4491	5292	-	STC1_Bam_061	135	123800	
6	7547	STC1_Bam_035_P16_M13.f	1	4	776	138	30580	8144	8919	-	STC1_Bam_035	135	127426	
7	7549	STC1_Bam_061_I18_M13.r	1	4	340	138	30580	8907	9246	+	STC1_Bam_061	139	106114	
8	7557	STC1_Bam_049_L03_M13.f	1	4	792	138	30580	13390	14181	-	STC1_Bam_049	135	132469	
9	7568	STC1_Bam_062_A13_M13.f	1	4	826	138	30580	20975	21800	+	STC1_Bam_062	8899		

1. Wiemy który contig leży na którym chromosomie (Ctg_ID_NCBI i Polskie chromosomy w pliku STC_vs_contigs1).
2. Nie znamy długości chromosomów.
3. Nie wiemy gdzie contigi leżą na chromosomach – pozycje start i stop (plik links3) są bezwzględne, nie dzielą organizmu na chromosomy.
4. Jak powiązać Ctg_ID_NCBI z contigami csb10a_v1 ?
5. Nie znamy sekwencji contigów identyfikowanych Ctg_ID_NCBI
6. Nie znamy sekwencji BAMów
7. Znamy położenie BAMów względem contigów (o id Ctg_ID_NCBI) – od 1st_base_ctg do Last_base_ctg (plik STC_vs_contigs1)